June 11, 2025

Dear Amanda,

Our submission was originally considered for publication and rejected at the *AER*. We are using the forwarding service to attach our reviews from this process. We have since revised the paper, taking into account the suggestions from the reviewers. This document details our responses to reviewers.

In this document we will start by providing a short summary of the changes we made. We will then provide detailed responses to the reports from each of the reviewers individually. As you will notice, this document ended up being quite long. To make it easier to navigate, click the hyperlinks below to be taken to a specific section of the document. The hyperlinks are also embedded in the header of the subsequent pages. In our experience as reviewers, we have noticed that the AEA journal submission formatting system sometimes destroys hyperlinks in PDF files; if this happens, you can find another version of this document at https://zarekcb.github.io/MAvsFFS_AEJ_ResponseLetter.pdf.

There are two pieces of important context for you and the reviewers that we wanted to put forth as you consider this submission, that we were unsure whether/how to discuss in the prior submission. First, the data access for this project is scheduled to cease at the end of September, giving us a shorter window to do revisions than would be typical (we have also had some intermittent lapses in access, leading to delays in resubmission). Second, we are not able to expand upon the analytic sample described in the paper. We received it as a pre-prepared data product from Inovalon and they are not willing to generate another pull for us. We mention this not to get any special exemption from proper criticism, but just to put into context what we can and cannot change relative to what is in the current manuscript.

Thank you again for considering the paper for publication in *AEJ: Economic Policy*. We are excited about this revised version and look forward to hearing your thoughts on it. Please let us know if there are any questions or concerns in your consideration of this resubmission.

Best regards,

Zarek Brot, Yalun Su, Boris Vabson, Scott Bilder, Barton Jones, Iman Mohammadi, Zulkarnain Pulungan, and Christie Teigland

Summary of Changes

Major Changes to Exposition.

- We edited the abstract to include the fiscal cost estimates both including and excluding the costs of supplemental benefits (R2.5)
- We have evened out our discussion of selection bias, in line with comments by R1.2 and R1.3 and some feedback we got between submissions. For instance, we now summarize our data contribution a bit more modestly as:

  > *This dataset provides us with one additional set of observables previously unavailable to researchers: health care utilization just before individuals qualify for Medicare.*

  and have made it more clear that we cannot remove *all* potential sources of bias.
- We have expanded our discussion of our sample construction and its comparison to other samples, as requested by R1.1, R2.3, R2.6b, R2.6e R3.2, and R4.1. We now go into more detail about how our sample size changes as we restrict the sample, and we compare our sample to a broader sample of new 65-year-old Medicare beneficiaries from the MBSF. One thing we should note is that our sample is different: It is whiter and in higher-income areas. We have been up-front about this when it arises, as well as in the conclusion. We also discuss the implications of the fact that our identification strategy relies primarily on 65-year-olds for our estimates, as highlighted by R1.1.

New Exhibits.

- Appendix Figure A1 has been simplified to plot the geographic distribution of our sample, without the comparison to the MBSF.
- Appendix Table A1 is a new exhibit that provides a waterfall table of how sample sizes shrink as we add additional sample restrictions (R2.6e, R3.2, R4.1)
- Appendix Table A2 is a new exhibit that compares our sample to the universe of 65-year-old Medicare enrollees from the MBSF in MA/FFS (R2.3, R3.2, R4.1)
- Appendix Table A3 is a new set of estimates where we use Poisson regression to consider our estimates under the alternative assumption of parallel multiplicative trends, i.e. parallel trends in ratios (R3.1)
- Appendix Table A4 is a new set of estimates of our main regression comparing our primary matching approach to an approach where we pick matched controls without replacement (R3.3)

- We have reviewed *all* exhibit notes and tried to make them more descriptive where they were previously lacking. (R2.6f)
- Figure 6 no longer presents results on readmissions since we no longer have faith in those results (see our response to R1.M1)
- Fiscal cost decompositions now have relabeled "Plan Bidding" as "Residual" to correctly reflect that this is a residual term (R1.4)

Reviewer 1

(R1.1) *Limited nature of the panel - The primary analysis focuses on individuals aged 65 and therefore represents the effect of just one year of MA enrollment. For context, the average age of MA and FFS enrollees in 2019 was approximately 71, and more than 60% of MA enrollees are age 70 or higher.1 In a supplementary analysis the authors use a longer panel of 2 years to examine the effects over ages 65-66 and estimate larger decreases in spending of 8-12%, depending on the length of the pre-Medicare period they observe in their sample. A caveat here is that the sample sizes for longer panels differ considerably. The larger savings estimate raises the question of whether the effect of MA enrollment increases in magnitude as people age. Conceptually, I expect the savings effect of managed care to grow with age. Beneficiaries typically stick with MA plans for a long period, unlike employer insurance and Medicaid. Since health is a stock, early investments could plausibly produce large gains later. MA plans therefore have an incentive to invest in preventive care early to reap savings over time as their members age and need more costly care. Other features of managed care, such as directing enrollees to higher quality providers and other utilization restrictions, could also generate larger gains when people use more care. How would the savings estimate differ if we had data from an appropriately representative sample of MA beneficiaries? At the same time, it is possible that the effects on health may also differ for older people. While restrictions on hospital use at younger ages do not lead to adverse consequences, this may change as people become frailer. Hence, not only the effect on spending, but also interpretation of the reduction in spending may differ between relatively younger and older beneficiaries.*

While this is an excellent comment, we don't think we have any reasonable way of addressing it throguh analysis. In the end, it is true that 65-year-olds are different from the general Medicare population. We considered trying to reweight to the general population on illness, but there are serious common support problems (most of our sample have very few chronic illnesses measured at 64), and so we would have to lean very heavily on a small number of beneficiaries in our sample.

We have added a line to our intro (see R1.2) and changed our final paragraph to bring this discussion more to the forefront (new part bolded):

> *Our estimates are suggestive of the potential effects of such sweeping changes to MA plan design and reimbursement regulation. However, we have two important limitations.* ***First, both our data and identification strategy impose sharp restrictions on which beneficiaries we can estimate effects for. Our identification strategy requires us to only estimate treatment effects for the youngest Medicare beneficiaries (who likely are lower-risk than older beneficiaries); our sample is whiter and higher-income than average new beneficiaries.***

> ***Our results suggest that these characteristics are all associated with smaller treatment effects, suggesting that our results may be smaller than the true average (selection-corrected) treatment effect of MA on utilization.*** *Second, we cannot fully assess the effects of regulation since we do not estimate how insurers respond to changes in MA policy (Duggan et al., 2016; Cabral et al., 2018; Miller et al., 2023; Vatter, 2024; Zahn, 2025). Evaluating the consequences of counterfactual payment and design reforms is an important area for further work.*

(R1.2) *Eliminating selection – The authors stress that one of the main contributions of this study is the ability to separate selection and treatment effects using baseline data prior to entry into Medicare. I am sympathetic to this claim since the difference in spending at age 64 between FFS and MA represents selection, assuming other factors are held constant (e.g., plan design). However, the research design relies on matching and cannot fully eliminate unobserved confounders. I am not sure if the design completely eliminates the possibility that some of the incremental difference between the two groups at age 65 also represents selection. MA plans expect to retain enrollees for many years and have an incentive to select unobservedly healthier folks not just at age 64 but also those predicted to have lower spending in the future. It is well known that medical spending grows nonlinearly among the elderly. It is therefore possible that the selection bias component also grows with age and imposing linearity, as done here, may not eliminate it entirely. This concern is highlighted by the fact that estimates of the effect of MA on spending from prior studies don't line up in ways we would expect. Schwartz et al. (2021) use the same research design as the present study, but a much smaller sample of Medicare beneficiaries, and find that MA reduces spending by 30% at age 65. In contrast, Curto et al. (2019) use a less robust design of comparing risk-adjusted spending for MA and TM enrollees in the same county and obtain estimates of 9-30% reduction in spending. Jung et al. (2024) use a similar design as Curto et al. but with more representative data and estimate a 12-18% reduction in spending. Note that both these studies included older people in their samples and that could reflect their higher savings estimates. Since studies using presumably weaker designs have found similar reductions in spending, it is difficult to claim that this study's estimates are free from selection bias. I suggest toning down the language around this claim.*

We have tried to tone down our discussion in the paper (especially in the introduction) that our results are purged of selection bias. We agree with you that our previous discussion was a bit too aggressive and one can never totally purge selection bias.

We have changed our discussion of the literature (emphasis here only):

> *We contribute to a long literature on the economics of the MA program. Prior researchers have struggled to find a way to account for the role of unobservable factors*

> *in driving selection into MA (Nicholas et al., 2024). Around 90% of studies in this literature rely on assuming away any differential selection into MA vs. FFS on factors not observable within standard CMS data to interpret their findings as causal (Agarwal et al., 2021; Ochieng and Fuglesten Biniek, 2022), though some studies use these observables in a more sophisticated way (Curto et al., 2019; Jung et al., forthcoming).* ***We show that substantial risk-adjusted differences between MA and FFS enrollees exist even before they enter Medicare, suggesting that the risk adjustment strategies used in this prior studies may not fully correct for selection bias.*** *Our approach is closer in spirit (and our estimates closer in magnitude) to other recent difference-in-differences approaches (Duggan et al., 2018; Schwartz et al., 2021, 2023). However, these studies have often been limited to small sample sizes and/or have only been measure a narrow set of outcomes, both problems that are solved by our new dataset.*
>
> *We also join a literature on the fiscal consequences of MA. Researchers have emphasized the role of favorable selection under imperfect risk adjustment in driving up the cost of Medicare (Newhouse et al., 2013, 2015; Brown et al., 2014). They have done this by focusing on beneficiaries who switch from FFS to MA, which is quite uncommon, leaving the potential that selection in this group may not reflect selection of inframarginal MA enrollees. We confirm the presence of favorable selection among initial enrollees. Our estimates of selection are slightly lower than many recent comparable switcher estimates (Jacobson et al., 2019; Lieberman et al., 2023), though not all (MedPAC, 2024). We find, echoing other recent work, that MA has resulted in increased fiscal costs of the Medicare program (MedPAC, 2025).* ***We estimate smaller fiscal effects of MA than prior studies (significantly smaller when accounting for differential plan generosity), though the differences in our estimates reflect both our ability to further correct for selection using pre-Medicare utilization and the fact that we are only able to employ this correction for 65-year-olds, rather than for all MA enrollees.***

Our goal was to emphasize our incremental contribution (the ability to control for additional unused observables) without mischaracterizing the literature. Note the last line also (we think) helps discuss our design-induced data limitations that you mention in R1.1. It also, in our view, helps explain the comparison in your last line here. We agree that some other estimates are of similar magnitudes, but the differences are both in research design *and* sample composition. Our approach allows us to disentangle the contribution of research design, though not, of course, sample composition.

We also now make this note when we discuss research design:

> *The extent to which selection bias remains results in our estimates being either upper or lower bounds of the true causal effect of MA enrollment, depending on the sign of the residual selection. To the extent that pre-Medicare utilization is only partially able to control for favorable selection (i.e., if MA plans are still favorably selected even conditional on this utilization), our estimates will be upper bounds, and the true effect of MA will be smaller.*

(R1.3) *Contributions relative to prior work – Related to the previous point, the authors unfortunately have to grapple with the issue that a lot of work has been done on this topic already. We therefore have several estimates of the effect of MA on utilization and spending to choose from. Agarwal et al. (2021) do a systematic review of the prior literature. There seems to be a consensus now that MA leads to a reduction in hospital inpatient and nursing home use, and these are the primary channels driving the reduction in spending. Consistent with the present study, prior work has also found mixed evidence on quality. Jung et al. (2024) have shown that the reduction in spending is concentrated among those in HMO plans, although they also find a reduction among PPO enrollees. Although this latter study has a weaker design than the present work, it has the relative strength that it uses the universe of encounters for all MA enrollees nationally. It is possible that it estimates a higher rate of savings because the sample includes older enrollees. My sense is the present study needs to do more, perhaps on quality effects and mechanisms underlying the savings, to unambiguously make a contribution worthy of a top-tier journal.*

We think of a big part of our contribution as being able to cover some previously-covered ground with 1) somewhat better identification, though subject to the caveats above; and 2) all within the same dataset. We think it is quite important that we find that (for the same sample of beneficiaries) HMOs both have larger treatment effects than PPOs *and* they are more selected, to make the point we (now) make in the intro:

> *Moreover, although we find that favorable selection increases MA reimbursement rates in general, across plans, favorable selection is not necessarily an indicator of fiscal consequences. While selection is primarily driven by HMOs, they also have less deleterious fiscal consequences than PPOs. This arises from the fact that MA plans must fund the supplemental benefits that attract favorably-selected beneficiaries through rebates, which they only receive if they reduce utilization. This creates a perverse positive correlation between a plan's ability to achieve the goals of the MA program and the extent to which it earns excess payments from imperfect risk adjustment.*

Of course, what makes a significant contribution is in the eye of the beholder, but we hope you agree with our view on this!

(R1.4) *Analysis of fiscal spending consequences – I enjoyed this analysis and think it helps to put the results on utilization and spending into a larger context. Overall, it makes sense that MA costs the government more than TM because imperfect risk adjustment does not fully eliminate selection. This is a well-known problem with how the government sets rates for MA plans and is consistent with commentary in academic work and popular media. However, much of this analysis is descriptive and tends to rely on inputs other than the causal effect estimates. So, it feels a bit disconnected from the main analysis. It is also clunky. For example, the authors value supplemental benefits using the rebate amount. But Cabral et al. (2018) find that only about 50% of every dollar given to MA plans is spent on benefits for members. What is currently allocated to supplemental benefits therefore also includes profits for MA plans. The "plan bidding" component seems like a residual and a euphemistic term for profits and is uncomfortably large for PPO plans. This analysis also ignores other changes that would ensue if MA were eliminated. For example, Kate Baicker and co-authors have shown that markets with greater MA penetration have lower spending among TM enrollees (Baicker et al., 2013). For these reasons, I suggest reducing the prominence of this analysis if it is to continue in its present form.*

There are a couple of important points in here which we will try to address in turn.

On "plan bidding:" we think your point makes sense and so have relabeled it as "residual" because we agree with you that we were unjustifiably 'labeling the residual' (so to speak). We now discuss it more prominently (we hope):

> *When taking out all of these components, the final component reflects the residual gap between MA's private cost reductions and what it passes through to the government through the bidding system. This residual reflects three factors. First, plan bids may be greater than expected costs, to pay off their fixed costs or to earn markups when they face imperfect competition. Second, plans set bids uniformly across their entire enrollee population; the residual we estimate is only for 65-year-olds, but may be of different size or sign for other groups. Finally, any measurement error in estimating the other components will feed into this residual.*

And discuss the HMO/PPO results in this context:

> *Though PPO plans do not appear to be favorably selected at all, they raise fiscal costs substantially. We estimate this largely comes from the fact that PPOs (1) raise utilization somewhat, rather than lowering it; and (2) they have larger residual components. This latter effect suggests that PPOs have higher markups than HMOs, though this only applies to 65-year-olds. This might arise even if the two types don't have strictly different overall markups if, for instance, PPOs enroll older populations and risk adjustment fails to fully correct for this, thus influencing plan bids (Orsini and Tebaldi,*

> *2017).*

On valuing rebates: This is part of a general class of issues with thinking about GE effects of some counterfactual plan design. This is easier when the switch is between status quo MA and moving everyone back into FFS (net of the spillover effects you mention), but indeed hairier when we think about what we call "standard coverage." We discuss it here (new part bolded):

> *To make such an apples-to-apples comparison, we ideally would want to break out the cost of providing these services. We do not, however, have accounting data that allows us to do this. Instead, we use the fact that these services must be financed out of the rebate payments that plans receive, rather than the base payments, and therefore proxy for the differential cost of different benefits offerings with the size of the rebate. We include both the core rebate amount here, as well as any rebate payments received from the Quality Bonus Payment program.* ***Note that this approach is a partial equilibrium approach—to interpret it as the counterfactual amount that CMS would necessarily spend if MA plan design was restricted, we would also need to assess how bids and selection would respond, which we cannot. In the absence of these rebate payments, benefits may not fall dollar-for-dollar (Duggan et al., 2016; Cabral et al., 2018). Moreover, as we show below, selection is quite different between plans that offer more vs. less supplemental benefits, meaning the group of MA enrollees might change compositionally.***

On spillover effects of MA, we have added in a footnote:

> *Note that, in this section, we are assuming that the only influence of the MA program is on MA enrollees. For instance, Baicker et al. (2013) find positive spillovers from MA onto FFS enrollees, and we will undercount these spillovers. Similarly, we will undervalue any extent to which the MA program puts pressure on providers or generates insurer participation in other related insurance markets.*

(R1.M1) *The effect on readmissions seems implausibly high. I would probe this further. Prior work has not found such large reductions.*

We have taken it out. Basically, the issue is that we had defined readmissions unconditionally (i.e., there were beneficiaries without index hospitalizations in the denominator). If you restrict the denominator to only those with index hospitalizations, the sign of the effect flips. The problem, though, is compositional: MA also reduces hospitalizations overall, including index hospitalizations. If the marginal *index* hospitalization avoided by MA has a lower risk of readmission (which is reasonable to assume), then when we condition on index hospitalization, hospitalized MA enrollees are, on average, *more* likely to be readmitted, as we find. It is hard for us to build a predictive model of readmission

since the count is not large in our data, so rather than engage in a complex procedure to correct for this in a tangential result, we thought it more prudent to just drop it.

(R1.M2) *Since the Inovalon data is a nonrandom sample of MA enrollees, the authors could consider probing the generalizability of their estimates by reweighting the sample to match the population share of MA enrollees across different states and/or demographic groups.*

Most of the places where reweighting would be the most useful–states, risk score, possibly race, and especially age–are areas under which we are woefully underpowered (or, in the case of states and age, simply do not have data), so we would have to lean very heavily on a very small part of the data.

Reviewer 2

(R2.1) *This paper is a useful application of a novel dataset and generates results with policy relevance. However, the paper would be stronger with greater attention to the economics of the public-private partnership and how these results inform public economics in general. We cannot conclude from this paper that welfare would be increased if we eliminate MA (the CF in Section 4) because the welfare impact of MA is not calculated. And we cannot conclude that welfare would be increased if we reduce MA capitation payments because that would affect MA benefit design, MA enrollment, and consumer surplus. I – someone very interested in the MA program – am very excited about a new method of calculating the magnitude of selection not accounted for by the MA risk adjustment system, but other Top 5 readers may want to see broader implications. And to the extent that the paper is pitched as a new method of calculating the magnitude of selection, it would be helpful to more directly compare your results to previous estimates. (i.e., the Introduction describes previous work and then says, "we confirm the presence of favorable selection" but the magnitudes are not compared.)*

Thank you for your kind comments! As for the last line (which is the most actionable) we now include this in two parts.

In the intro:

> *Our estimates of selection are slightly lower than many recent comparable switcher estimates (Jacobson et al., 2019; Lieberman et al., 2023), though not all (MedPAC, 2024).*

When the estimates are presented:

> *Our estimates **[NB: 11.1%]** are smaller than Brown et al. (2014), who estimate that MA enrollees are 16-22% less costly, and larger than Newhouse et al. (2015), whose estimate is 5%. These studies both study selection in the late 2000s. Our estimates are also smaller than recent white paper estimates which cover more time periods that overlap with ours, such as Jacobson et al. (2019) (13%) and Lieberman et al. (2023) (21.5%), though not recent estimates by MedPAC (2024) (4-6%). All of these studies focus on those who switch from FFS to MA, not those who enroll in MA upon qualifying for Medicare.*

(R2.2) *I appreciated that the authors are straightforward about the need to assume that utilization at 64 is experienced under some kind of uniform ESHI, or in particular that any aspect of ESHI that causally affects 64yo utilization (e.g., ESHI HMO) is independent of MA enrollment. There's been very little data about the ESHI-Medicare transition b/c as far as I know we were previously limited to small samples (Health and Retirement Survey plus a few*

Federal surveys with a short panel). That makes it hard to evaluate this assumption.

The paper would be strengthened by a greater attempt to explore this key identifying assumption. Is there any way to apply the Inovalon data to examine this pattern? For example, does ESHI HMO share at 64 in a state predict MA share at 65? I am also quite interested to know the persistence of parent insurer across this transition, which could be quite strong if ESHI providers advertise (perhaps selectively!) to their 64yo members. This persistence would violate the exclusion restriction because it will mean that MA is likely particularly strong in, e.g., states with high UnitedHealthcare shares. (Alternatively, if you believe your matching addresses this, please add material explaining how.)

I at first had concerns about the paper's empirical design in light of the great variation in geographic uptake in MA. However, I eventually realized that your matching estimator matches on state X urban, which I believe would account for a lot of this geographic variation. I would prefer matching within counties, although I would be fine with year or quarter of birth (instead of month) if needed to increase cell size.

Thanks! We are a little confused by the HMO share comment—as we highlight in our summary stats table, the at-64 HMO share is higher among MA enrollees, so mustn't it necessarily be that the converse is true? We reviewed this comment a few times and weren't sure what you were thinking.

On the exclusion restriction comment: Please let us know if we are misunderstanding, but we don't think this is a threat to identification. In general, the set of insurers offering MA policies is a *superset* of those offering ESHI policies. Generally, ESHI requires scale and thus only larger insurers participate, whereas MA includes some smaller insurers. Therefore, any 65-year-old should always have an option to continue with their same insurer, and this shouldn't vary (much) across regions.

On some of the data requests: For the carrier questions, we are happy to answer in this response letter but our data use agreement with Inovalon prevents us from using the information on the carriers in the manuscript (due to their worries about carrier identity disclosure). In our analytic sample, 75% of MA enrollees have the same MA carrier as the one who provided their ESHI plan at age 64. However, this is partially mechanical: Since we sample at the insurer level, and individuals get thrown out of our data if we cannot see them at both 64 and 65, we are throwing out many of those who enter into an MA plan not tracked by Inovalon from an ESHI plan tracked by Inovalon.[1] On selection, we initially had the same thought as you. However, same-insurer enrollees are *less* selected than different-insurer enrollees. The sampling may be responsible for this; but also, as Shapiro (*Marketing Science* 2018) estimates, advertising is not very effective as a cream-skimming device. (Note that Aizawa and Kim 2018 find the opposite in MA, but primarily in the era before HCC-based risk adjustment; in

[1] If you include those within our summary stats table in the "Non-Inovalon" column, who we cannot track post-65 utilization for and must always be different insurers, the rate drops to 24%.

their counterfactual simulations, they find that HCC risk adjustment reduces the selection benefits of advertising substantially)

We agree that matching within county might be ideal. However, our final analytic dataset does not actually include county identifiers, so we are not able to do this. We now make this clear in a note in the data section:

> *Due to privacy concerns, the final analytic dataset does not include identifiers (not even anonymized) for counties or plans. To preclude potential re-identification of individual counties and plans, we are only able to observe ventile intervals of bid, rebate, and benchmark amounts; when we use those variables for each county/plan we substitute the mean value of the associated ventile.*

(R2.3) *An unfortunate weakness of the Inovalon data is that it is a convenience sample of participating insurers who agree to supply claims (or, as it is correctly stated on page 9, "determined by which insurer clients they have contracted with.") I was therefore surprised to read it described in the Introduction as "nationally representative". In the Data section, these claims are somewhat weaker: "a significant share of the nation's insurers". A map is provided in the Appendix to support the claim that it is "broadly representative"; however, that map only informs me about state representation (itself not great) and not within-state geographic representation (e.g., urban-rural), or any nongeographic dimension such as health risk. I encourage the authors to be forthcoming about the weaknesses of the data so that they can more honestly discuss how to interpret their results in light of them.*

Thank you, we agree that we were a bit too strident in our earlier draft. We have removed "nationally representative" from the intro, and in the Data section, we now have

> *The extent to which our sample is broadly representative depends on the geographic coverage of the Inovalon data. This is determined by which insurer clients they have contracted with. In Appendix Figure A1, we plot the relative geographic distribution of our analytic sample. We are able to track MA enrollees in 32 states. Our MA coverage is especially dense in the Northeast and Great Lakes regions of the US. On the other hand, we have very little coverage in the Great Plains and Deep South. Our FFS coverage is closer to proportional relative to state populations. A more substantive limitation of our dataset (motivated by our empirical approach, described in the next section) is that we only observe new 65-year-old Medicare beneficiaries, who are young relative to the average beneficiary who qualifies by age.*

We also have included, in Table 1, a comparison of our sample to the universe of 65-year-old MA/FFS enrollees. We discuss it:

> *How does our sample compare to a broader population? We make two comparisons. First, we can compare MA enrollees in our analytic sample to MA enrollees who we can track at age 64, when they are enrolled in an Inovalon client ESHI plan, but not age 65, since they enroll in MA plans that are not Inovalon clients. This group is largely similar to MA enrollees in our analytic sample, except they are somewhat less likely to be white (84% compared to 91% in our analytic sample) and have lower average quarterly spending at age 64 ($1,175 vs. $1,464). We display this comparison in Table 1.* ***Second, we can compare to a broader population of 65-year-old MA/FFS enrollees represented in the Medicare Beneficiary Summary File. We cannot link this group to Inovalon data, meaning we do not have fine-grained socioeconomic data, and must instead use characteristics at the county level rather than the 9-digit zip code level. Largely due to our limited geographic coverage, beneficiaries in our dataset are much more likely to be white, and live in areas with higher average income and higher home ownership rates. We display this comparison in Appendix Table A2.***

We also come back to this in the final paragraph:

> *Our estimates are suggestive of the potential effects of such sweeping changes to MA plan design and reimbursement regulation. However, we have two important limitations.* ***First, both our data and identification strategy impose sharp restrictions on which beneficiaries we can estimate effects for. Our identification strategy requires us to only estimate treatment effects for the youngest Medicare beneficiaries (who likely are lower-risk than older beneficiaries); our sample is whiter and higher-income than average new beneficiaries. Our results suggest that these characteristics are all associated with smaller treatment effects, suggesting that our results may be smaller than the true average (selection-corrected) treatment effect of MA on utilization.*** *Second, we cannot fully assess the effects of regulation since we do not estimate how insurers respond to changes in MA policy (Duggan et al., 2016; Cabral et al., 2018; Miller et al., 2023; Vatter, 2024; Zahn, 2025). Evaluating the consequences of counterfactual payment and design reforms is an important area for further work.*

(R2.4) *The authors describe large differences between MA HMOs and MA PPOs. I am not aware of a lot of papers that find major differences between HMOs and PPOs in MA, so this paper breaks new ground in this area. In addition, these results can offer CMS greater insight into the performance of the MA model. However, I felt that the authors misstated the regulatory differences between these two models.*

- *On page 2, PPO plans are described as having "no utilization management", even though*

PPO plans commonly use prior authorization to establish that care is medically necessary – KFF has a brief on this issue that states that 99% of MA enrollees have prior authorization for some services. Perhaps the authors are using the words "prior authorization" in an unfamiliar way to reference HMOs use of gatekeeping?

- *In addition, about one third of HMO enrollees are in HMOPOS plans that act like HMOs for certain services, PPOs for others (I assume HMO POS plans are included in the HMO sample, but perhaps I am wrong about that). This further muddies the distinction between these two types of plans.*
- *Institutional Background paragraph on page 7 that goes into greater detail on the HMO-PPO distinction has no citations, which might have corrected me if I am wrong.*
- *Why do MA PPOs appear to causally increase utilization? PPOs differ from FFS in two ways that would tend to reduce utilization: encouraging in-network provision (should weakly reduce prices and utilization) and requiring more prior authorization than FFS (e.g., see the KFF piece). If I accept the core identifying assumption of the paper, and accept the methods used to account for health risk, what explains this result?*

Thank you for these comments. We were a bit too casual in our initial discussion and have tried to tighten things up in line with your suggestions.

On PPO prior authorization: You are right that there is some prior authorization among PPOs, so we have replaced the referenced sentence within the introduction:

> *We compare HMO plans, which impose stronger restrictions on beneficiaries, to PPO plans, which impose weaker restrictions.*

In general, the mandated PCP gatekeeping necessitates more use of prior authorization in HMOs than in PPOs. We have changed the paragraph within the plan design part of the background section. We also reference your (correct) guess that we do categorize POS plans as HMOs:

> *HMO plans differ from PPO plans by using additional utilization management tools. HMO plans mandate care coordination through the patient's (explicitly-designated) primary care provider, meaning that enrollees must obtain a referral before seeking specialty care. By virtue of this gatekeeping, HMOs enforce greater utilization management through policies like prior authorization than PPOs do. PPOs typically provide some limited coverage for services rendered by out-of-network providers, whereas many HMOs do not provide any such coverage.*[a]
>
> [a] Point-of-service or HMOPOS plans (which we categorize as HMOs) do, however, provide some out-of-network coverage and occasionally allow the patient to bypass their primary care provider.

On why PPOs increase utilization: Our view is in this line we have now added to the HMO/PPO heterogeneity results:

> *To the extent that PPOs increase utilization despite having some use of managed care, this is likely to come from the fact that cost-sharing faced by MA enrollees is, on average, more generous than FFS enrollees.*

(R2.5) *I read Section 4 twice and I still found it difficult to understand the methods. Perhaps you can design a set of equations or some kind of exhibit that helps explain your thinking? Here are some questions/comments:*

- *A distinction is made between upcoding and selection that I do not quite understand. One possibility is that the upcoding element is a percentage adjustment taken from previous work, while the "selection" adjustment is this paper's original analysis and leverages the age-64 spending. In Figures 7 & 8a, the selection adjustment is much larger than the upcoding adjustment. Does that imply that, under the identifying assumption of your paper that age 64 utilization is a "ground truth" from which we can infer CF age-65 spending, you find much more selection into MA than we previously understood? (If so, is this not a finding you'd like to highlight in the Introduction?)*
- *What is the plan bidding element? Is this the residual after your parametric adjustment for the other elements? I did not understand how you calculated this item, and that made me very uncertain of its interpretation.*

That said, once I understood how you are accounting for the rebate, I object to the use of the 15% statistic in the abstract, and believe the 3.6% is a truer statistic to report there. The 11.4% that contributes to lower cost-sharing, extra benefits, and extra subsidies to Part D prescription drug insurance is expenditure with real value to consumers, whereas the selection component is straightforward rent-seeking.

This comment contains a few points that we will address bit by bit.

Upcoding vs. selection: Here the distinction is that selection is the extent to which *differentially profitable* beneficiaries enroll in MA, whereas upcoding is about actions MA plans take *after beneficiaries enroll* to make them more profitable (by adding extra diagnoses). We do not find significantly more selection than other studies, though we now mention this comparison in the paper (in response to your comment R2.1).

Plan bidding: See also our response to R1.4. The reviewer correctly pointed out that this is us labeling (possibly mislabeling) a residual. We now discuss it differently when it is introduced:

> *When taking out all of these components, the final component reflects the residual gap between MA's private cost reductions and what it passes through to the government through the bidding system. This residual reflects three factors. First, plan bids may be greater than expected costs, to pay off their fixed costs or to earn markups when they face imperfect competition. Second, plans set bids uniformly across their entire enrollee population; the residual we estimate is only for 65-year-olds, but may be of different size or sign for other groups. Finally, any measurement error in estimating the other components will feed into this residual.*

As well as its role in explaining the HMO vs. PPO difference:

> *Though PPO plans do not appear to be favorably selected at all, they raise fiscal costs substantially. We estimate this largely comes from the fact that PPOs (1) raise utilization somewhat, rather than lowering it; and (2) they have larger residual components. This latter effect suggests that PPOs have higher markups than HMOs, though this only applies to 65-year-olds. This might arise even if the two types don't have strictly different markups in a global sense if, for instance, PPOs enroll older populations and risk adjustment fails to fully correct for this, thus influencing plan bids (Orsini and Tebaldi, 2017).*

On 15% vs. 3.6%: We have sympathy for your view on this, and we weren't sure what the "right" estimate to put in was, so we tried to split the difference and make room to include both. As we discuss in the conclusion, without the supplemental benefits, it is hard to imagine that beneficiaries would have any desire to enroll in MA, so the question is whether you want to evaluate the MA program as a whole or the specific potential role of private provision relative to public provision. We think both are quite important.

(R2.6a) *"For each measure, we sort plans and counties into twenty ventiles bins, and then calculate the mean values within each bin." Is this just for some of the descriptives (which?) or is it for parts of analysis? For the fiscal calculations?*

Hopefully this is clearer from the quote we mentioned in R2.2. This is due to privacy restrictions to prevent us from reidentifying counties/plans, so this is used for every piece of analysis.

(R2.6b) *"We additionally merge in public CMS data on plan-level MA bid and rebate amounts", but FN 7 indicates you do not observe the names of the insurance carriers or their plans. If you can't get the names, I can't imagine they would give you the CMS contract IDs. How do you accomplish this merge? No details of the "public CMS data" are given; is it the 422.272 bid pricing data?*

Thanks for this question. We realized our prior discussion was a bit unclear. To summarize for you: Inovalon observes the contract IDs in the raw data, but we don't in the final data. We have them merge in the bids in the raw data, then they anonymize them when they prepare the sample for us. We now include this as a footnote:

> *Due to privacy restrictions, the final dataset we use does not include the names of the insurance carriers, nor those of their plans. Inovalon allowed us to link plan characteristics (described below) before the plan de-identification process.*

This is indeed the 422.272 bid pricing data.

(R2.6c) *Finally, it appears you calculate risk scores at age 64, which means you apply the HCC model to commercial claims (or is it a "risk adjustment factor" that is not an HCC that is occasionally mentioned?). It would be helpful if you can find any validation of the HCC score calculated on commercial claims; I am uncertain of the comprehensiveness of diagnostic coding in commercial claims.*

You are correct that we are applying the HCC model to commercial claims (we hadn't realized that wasn't stated explicitly!). Our guess is that you are worried that we will fail to accurately capture the risk score, and thus our measure of selection is too large due to the fact that the risk score used by CMS is richer than ours. We respond to that in new text in that section:

> *We construct our best estimate of the CMS-HCC risk score used to adjust payments by constructing it based on diagnoses observed at age 64.*[a] *If risk adjustment perfectly captures selection, our estimates should shrink to zero when we control for the risk score.*
>
> [a] The employer-sponsored claims data tend to report less diagnoses per claim line than FFS Medicare data, so we may be able to observe less diagnoses than CMS is normally able to. This will not bias our estimates of selection *unless* selection is correlated with conditions that are differently recorded between the two sets of claims data.

We of course cannot test whether selection is correlated with differential recording, since post-65 coding is quite different in MA vs. FFS both due to data reasons and insurer strategy.

(R2.6d) *What is "base payment benchmark adjustment" under Quality Bonus Program in Table 3?*

As part of our response to R2.6f we have now made our exhibit notes much more descriptive, so we can copy from the new note:

> *"Quality Bonus Payment" reflects additional payments based on the quality bonus payment system: increases to rebate-sharing rates ("Rebate Multiplier") and increases to county benchmarks that both increase rebates ("Rebate Benchmark Adjustment") and base payments ("Base Payment Benchmark Adjustment").*

To be a bit more verbose: The QBP program increases the benchmark that plans face. All else equal (including bids), that raises their base payments. (Note that, if the plans were bidding above the benchmark, some of this money is effectively transferred from CMS to beneficiaries rather than to plans, but bidding above the benchmark is infrequent)

(R2.6e) *While I recognize the sample design is not straightforward, an N would still be useful. How many MA enrollees with observed 64yo utilization are successfully matched to FFS enrollees with observed 64yo utilization, and how representative are they of the overall MA population on geography, income, and risk score? In addition, how many Inovalon-sampled MA enrollees were excluded due to (1) failure to observe 64yo utilization or (2) failure to match to an FFS enrollee with observed 64yo utilization?*

Thanks for this comment, which echoes similar comments by R3.2 and R4.1. One of our new exhibits is Appendix Table A1, which provides a waterfall table of how each data restriction imposed reduces our sample size. Unfortunately, as we discuss in our cover letter, we cannot go back and evaluate summary statistics at each of these stages, since we do not have access to these "earlier" supersamples. The failure to observe utilization at 64 cuts our sample by about half. In contrast, failure to match only eliminates 10% of the sample (providing a levels comparison between the two is not informative since the first restriction occurs so much farther upstream in the pipeline). In Table 1, we now report a comparison to a broad sample of 65-year-olds in Medicare, which we think hopefully addresses your general concern.

(R2.6f) *As a general tip I would have appreciated more detail in the figure and table notes.*

Thank you. We have gone through the exhibits and tried to buff up the notes; if you have concerns about any specific exhibits in the revised submission, please let us know.

Reviewer 3

(R3.1) ***Selection on Outcomes****: one of the many nice features of the paper is that the authors are transparent about the assumptions underlying their empirical strategy. As they note, the main challenge in estimating the causal effect of MA is eliminating the selection effect. To clarify the following argument, I will refer to MA as the treatment and TM as control. The treatment-group corresponds to those who eventually receive the treatment and, equivalently, the control-group those who eventually obtain FFS coverage. The central challenge for the paper is that there is selection into treatment. The central assumption in the paper is that differences in treatment outcomes are independent of selection into treatment conditional on pre-treatment outcomes.*

This is not a minor assumption. If pre-treatment outcomes were always sufficient to control for selection into treatment, then we would not need randomized control trials. However, we know that in practice, individuals select into treatment in ways that express differences in their expected outcomes. In the case of Medicare, those who select into FFS (+Medigap) are likely those who expect to expand their utilization more under more generous coverage than those who select into MA. Differences in pretreatment periods cannot account for this, as the pre-period does not express treatment differentials. More concretely, the pre-period used in this article does not express the intensive margin adjustments to changes in coverage (i.e., moral hazard). If the authors had richer detail on ESHI coverage and variation in it, then they might be able to estimate pre-period moral hazard and then adjust post-period effects for it.

Given how strong this assumption is, I would have expected a stronger argument for it in the paper. In the notation of the paper, we must accept that $E[Y_i(FFS)|MA_i = 1] - E[Y_i(FFS)|MA_i = 0] = E[Y_i(ESHI)|MA_i = 1] - E[Y_i(ESHI)|MA_i = 0]$*. However, ESHI is often very different than FFS. If anything, it is likely more the case that* $E[Y_i(MA)|MA_i = 1] - E[Y_i(MA)|MA_i = 0] = E[Y_i(ESHI)|MA_i = 1] - E[Y_i(ESHI)|MA_i = 0]$*, as MA plans resemble the structure of standard ESHI plans. In fact, looking at the event studies in the paper, it seems that* $E[Y_i(MA)|MA_i = 1] \approx E[Y_i(ESHI)|MA_i = 1]$ *net of some overall age-trend. In general, if we posit that MA and ESHI coverage are not all that different, the author's assumption requires approximately that* $E[Y_i(FFS)|MA_i = 1] - E[Y_i(MA)|MA_i = 1] = E[Y_i(FFS)|MA_i = 0] - E[Y_i(MA)|MA_i = 0]$*, that is, homogenous treatment effects. This implies that anticipated differences in spending do not drive differences in TM/MA choices across individuals. This is despite large differences in coverage and access to care between MA and TM. This rules out selection on moral hazard and much of the basic theory on adverse selection and contract design.*

The argument above guides my interpretation the results: I see the evidence as more indicative of the effects of TM+Medigap relative to something similar to continued private coverage. Under this interpretation, the results in the article suggest a combination of strong selection effects and strong moral hazard effects from expanding coverage and access. The relevant mechanisms to explore in this interpretation, however, would be those associated with differences in Medigap coverage, and local access to FFS care, and potentially differences in pre-existing ESHI access. The current paper, however, does not speak to that.

We agree that this is important, and we appreciate your thoughts on this, which are more sophisticated than how we had originally thought about it.

You are right that, if anything, our time series could suggest that we are recovering the ATU, rather than the ATT. Reframing your concerns in a much more simplistic way, imagine that the true utilization-determining process is multiplicative in the underlying plan design attributes. Since FFS appears to increase utilization relative to ESHI, it might also widen the gap between MA and FFS enrollees relative to ESHI if both were to enroll in it.

We have tried to respond to this in a few ways. First, we rearranged our discussion of research design to highlight this potential issue, and give our response to it:

> *The largest threat to this assumption is the (implicit) functional form assumption that must be made: We must assume that level differences under ESHI are equivalent to counterfactual level differences under FFS, in order to unbiasedly estimate the ATT. If, for instance, utilization is multiplicative in arguments and (as we show later in this section) FFS raises utilization relative to ESHI, counterfactual level differences under FFS may be larger than observed level differences under ESHI. In this case, our estimates should be thought of as upper bounds on the true treatment effect of MA.*[a] *In contrast, we do not face any threat to identification from behaviors such as intertemporal substitution—i.e., individuals delaying care until they turn 65 to take advantage of more generous Medicare coverage—as long as this behavior is identical across future MA and FFS enrollees.*
>
> [a]Below, we show that utilization under ESHI is similar to utilization under MA. If this is the case, our coefficients will instead estimate the average treatment effect on the *untreated* (ATU). The ATU is equal to the ATT so long as there is no selection into MA based on its potential (individual-specific) treatment effects. Such an assumption is somewhat implausible given that we find later that a major selection margin is to avoid managed care, potentially implying that those who select into MA expect to be less affected by managed care. However, there are reports that some MA enrollees were unaware of what they were signing up for, sometimes being unaware that they were enrolling in MA at all (Abelson and Sanger-Katz, 2022), potentially blunting any such precise selection on treatment effects. In the next section, we estimate treatment effects under an alternative parallel trends assumption for robustness.

Second, we have tried an alternative identification approach at the end of Section 3.2:

> *We also explore robustness to alternative functional form assumptions. As we discuss in the prior section, we must assume that differences at age 64 in ESHI are equal to counterfactual differences in FFS. One reason that might not be true is that utilization is higher in FFS than in ESHI for FFS enrollees. If differences across individuals in spending are from factors that are multiplicatively separable rather than additively separable, differences at ESHI will be underestimates of selection bias. We therefore consider estimation of effects on utilization under an alternative assumption: that, if we forced both groups into FFS, the ratios of average spending between FFS and MA enrollees would be the same as they are when both groups are enrolled in ESHI at age 64; i.e., that* $\frac{E[Y_i(ESHI)|MA_i=1]}{E[Y_i(ESHI)|MA_i=0]} = \frac{E[Y_i(FFS)|MA_i=1]}{E[Y_i(FFS)|MA_i=0]}$. *We can then estimate multiplicative treatment effects as* $\frac{\frac{E[Y_i(MA)|MA_i=1]}{E[Y_i(FFS)|MA_i=0]}}{\frac{E[Y_i(ESHI)|MA_i=1]}{E[Y_i(ESHI)|MA_i=0]}}$ *using Poisson regression (Wooldridge, 1999, 2023), which are comparable to our estimates of percent changes in our main specification. We run specifications with and without using age 64 data, and with and without matching.*[a] *Our estimates (shown in Appendix Table A3) of the risk-adjusted difference under this specification are much larger (estimated reductions of 23.0% and 26.2% compared to OLS estimates of 17.7% and 19.3%), but our difference-in-differences specifications are comparable (estimated reductions of 9.6% and 6.3% compared to OLS estimates of 8.3% and 7.5%), suggesting that our specific parallel trends assumption is unlikely to have significant influence on our estimates.*
>
> [a]Note that we cannot run our primary specification since it would require estimating a maximum likelihood specification with many (beneficiary-level) fixed effects, and our server access did not allow us to add software to do so easily.

We think the argument that we are estimating the ATT is somewhat more plausible in ratios than in levels; we were reassured to find that the estimates are not very different (in percent change terms) when we do so.

To explain the footnote: While there are now packages to estimate Poisson regression with high-dimensional fixed effects, on the Inovalon servers (where the data lives) we only have access to SAS, which has no such package. This is why, for instance, we didn't re-run the entire analysis of the paper using Poisson regression, since we would have to devolve to a lesser specification.

(R3.2) ***Sample Selection and Data****: The information provided in the Inovalon data seems very helpful for this line of work. Notably, the pre-post data linkage is fantastic and a centerpiece of this article's contribution. However, I would have loved to see more details about how representative it is. Given the role of Inovalon in the market and the MA market share reported, I suspect it does not include information on some of the largest ESHI or MA insurers. I wonder if the evidence is thus representative of the average effects of MA (if we believe the estimates) or the effects on a subset of markets or certain demographics. I believe that some*

clarity on what is and is not represented would be enormously valuable, as the final matched samples seem quite small.

Another important caveat is the lack of allowed amounts or patient cost-sharing. The paper makes a case that MA spending is lower due to cost-control strategies, or equivalently, FFS increases spending due to the lack of such strategies (relative to ESHI). It seems important to know how much of the differential is passed through to consumers and how much insurers obtain. Past evidence cited in the paper suggests that insurers benefit greatly from this spending reduction.

Great suggestion. Three comments:

1. We have updated Appendix Figure A1 (the map of our coverage) to make it more readable, and updated our discussion in text:

> *The extent to which our sample is broadly representative depends on the geographic coverage of the Inovalon data. This is determined by which insurer clients they have contracted with. In Appendix Figure A1, we plot the relative geographic distribution of our analytic sample. We are able to track MA enrollees in 32 states. Our MA coverage is especially dense in the Northeast and Great Lakes regions of the US. On the other hand, we have very little coverage in the Great Plains and Deep South. Our FFS coverage is closer to proportional relative to state populations. A more substantive limitation of our dataset (motivated by our empirical approach, described in the next section) is that we only observe new 65-year-old Medicare beneficiaries, who are young relative to the average beneficiary who qualifies by age.*

2. We have updated Table 1 to include a comparison to a broad set of beneficiaries who are enrolled in FFS/MA at age 65. Reassuringly, our sample is very similar on observables, even if not totally on geography.

3. We have now included Appendix Table A1, which details how each of our sample restrictions reduces the sample size. We discuss it in the text (snippet below). One thing we should note is the awkwardness of having to bundle together the year restriction (2012-2019 $\rightarrow$ 2015-2018) with the 12 month restriction. Unfortunately, this table was generated in the data build process by Inovalon and we cannot go and disentangle it.

> *Between 2012-2019, about 1.3 million individuals were newly enrolled in the Medicare program at age 65, with 56% of them enrolled in MA. After our sample restrictions, our final sample composes of 205,557; 180,087 (87%) enrolling in FFS, and 25,470 (13%) enrolling in MA. Of the restrictions we impose, the ones that*

> *reduce our sample the most are a) the restriction to 2015-2018; b) the requirement for at least 12 months of coverage in the Inovalon data at age 64; c) no concurrent enrollment in commercial coverage at age 65; d) enrollment in a non-EGWP plan; and e) tracking of MA claims within the Inovalon dataset. In contrast, the requirements that enrollees stay in their plans continuously; that Inovalon tracks their commercial plans well; and that they enroll at age 65 do not bind as tightly. The final two requirements (that MA claims must be tracked, and that enrollees must be in non-EGWP plans) is especially binding for the MA component of our sample, resulting in MA underrepresentation within our sample; in contrast, we observe the universe of claims for FFS enrollees. We provide a step-by-step evolution of how our sample changes as we add restrictions in Appendix Table A1.*

On cost-sharing: While you are right that we cannot track cost-sharing within our data, we know from other data that cost-sharing is lower in MA than in base FFS (absent supplemental insurance). We reference this in the ppaer:

> *MA plans significantly reduce patient cost-sharing relative to the standard FFS arrangement; Ippolito et al. (2024) estimate that out-of-pocket costs are 18-24% lower in MA than in FFS.*

(R3.3) ***Empirical Strategy****: Conditional on the assumption discussed above, there are some details about the empirical strategy and evidence that I find worrisome. First, looking at Figure 2, the effect seems to be barely significant for a single quarter. For the sample that can be followed for two years, the effect on that quarter vanishes, and the impact is seen only in the last two quarters. I also find there is too much noise to conclude that the effect is monotonic, as suggested in the article. The article also should be clearer on how this two-year sample is matched, as it is not described.*

The strategy also uses a propensity-score matching that resamples controls. For individuals in the treatment group with few good controls, this will tend to provide a false sense of control stability, making the results artificially more significant. I would propose not resampling the controls or showing robustness to this. I also worry that matching based on how certain observables correlate with selection into MA might create somewhat artificial comparisons. Ideally, we would like to compare beneficiaries who are similar in risk and demographics and see how MA affects their utilization. By using a propensity match, we might be comparing two very different individuals who just happen to have a similar propensity to enroll in MA, albeit for very different reasons.

With respect, we think you are overindexing on the significance of each specific event study estimate, which (from our experience) is atypical in these sorts of DID approaches. Remember that quarter-by-quarter spending is much noisier than higher aggregates (e.g. year), which is why our preferred

specification for interpretation (Appendix Table B2) aggregates the estimate across quarters within a year.[2] Moreover, the second year increase is fairly clear to see in the raw data (Appendix Figure B1). Moreover, it is easier to see this increase over time in the 1 pre/3 post sample, where there is more precision (possibly due to the slightly larger sample size compared to the 2/2 sample). We do make a note of this in the relevant section though:

> *Though the estimates rise substantially, the standard errors do as well (especially for our sample with two years of pre- and post-enrollment data), and so we cannot always reject that estimates have not increased across post-enrollment years.*

We describe the matching for these subsamples in Appendix B (emphasis only here):

> *In our primary estimates, we focus on short panels analyzing four quarters before and after qualifying Medicare by turning 65. In this appendix, we extend our analysis to subsamples where we can observe beneficiaries for longer periods of time before and/or after turning 65.*
>
> *For each subsample, we construct it by starting with our initial sample, and conditioning on being able to observe continuous enrollment in either employer-sponsored or Medicare coverage for the target length of time.* ***For empirical analysis, when matching, we use the propensity scores constructed from the model estimated on the entire sample, but rematch treated beneficiaries to five control beneficiaries within the restricted subsample. We then truncate our observation window to the same period used for subsample selection.***

We tested robustness to your suggestion of matching without replacement. It has only a minor effect on our estimates. Here is our discussion of it:

> *We also estimate our results under alternative matching strategies. In the first, we match without replacement (in contrast with our primary approach, which matches with replacement) to avoid repeatedly sampling the same FFS beneficiaries. Our estimates (shown in Appendix Table A4) under this approach are only slightly larger (7.5% compared to 6.2% in our main specification).*

We are very sympathetic to your concerns about matching. In the end, our view is that adding matching at all does very little to the magnitude of our primary estimates (e.g. compare columns 2 and 3 or 6 and 7 in Table 2), so we view it as secondary in our empirical strategy relative to having the pre-65 utilization data. Our high-level results (differences between risk-adjusted vs. DID approach on utilization; differences between HMOs and PPOs; fiscal effects) are larger in magnitude than the changes that arise from tinkering with the matching approach.

[2] Note that we always cluster our standard errors at the person level so disaggregating down to quarter does not artificially boost our statistical power.

Reviewer 4

(R4.1) *Given that a key contribution of the paper is the novel data, I think it would be really valuable for the authors to show a bit more data to the readers. Currently, I find the data and sample descriptions in Section 2 to be quite opaque. For instance, in Section 2.1, the only statistic given about the dataset is that "the Inovalon data cover about 30% of the commercially-insured population." The authors then describe merging Inovalon with Medicare enrollment file and keeping the merged individuals, and going through various sample restrictions before arriving at the final sample for analysis. Understanding the data and the sample restrictions are especially important for ensuring representativeness of the sample, considering the modest size of the final MA sample (25,470 unmatched, or 22,784 matched, per Table 1). In this vein, I would suggest the following:*

a. *More information about the coverage of the Inovalon data. E.g. some information on the number of insurers and plans represented in the data would be valuable for thinking about both the data coverage as well as how heterogeneous the ESHI sample is, the latter important for interpreting the DiD results. (I understand that individual plans or names of insures cannot be disclosed per data restrictions).*

b. *To the extent possible, report the impact of various merges and sample restrictions on the sample size. Right now I find it hard to reconcile Inovalon being "30% of the commercially-insured population" with the modest final MA sample size (1% of total MA population). Which sample restrictions are responsible for such large discrepancies?*

c. *I find Appendix Figure A1 hard to parse. Based on the notes, the figure plots "the ratio of the state share of newly enrolled MA (or FFS) enrollees at age 65 based on our analytical sample divided by the share of newly enrolled MA (or FFS) enrollees at age 65 based on the Master Beneficiary Summary File (MBSF) base." Would it be possible to simply plot the share of the MA (FFS) population in the analytical sample? Ultimately, I think the readers are interested in the coverage and representativeness of the sample and this simple metric would be much more informative.*

d. *What data/sample does the last column of Table 1 (Non-Inovalon MA) come from?*

Thank you for these comments; this sentiment was shared by two of the other reviewers (R2.3, R3.2). One note on the small size of our final MA sample: Unfortunately, this is where the insurer-level sampling really poses an issue for us, since we lose MA enrollees much more drastically than FFS enrollees–we retain about 25% of our initial FFS population (including unused years in the denominator) but only 4% of our initial MA population.

a. We now mention in the main paper:

> *Our final dataset covers a little over 50 unique carriers both in ESHI and MA.*

The numbers are slightly different for ESHI vs. MA. Our DUA restricts us from being able to publicly disclose the exact number of insurers.

b. We have now included Appendix Table A1 to address this comment, which details how each of our sample restrictions reduces the sample size. We discuss it in the text (snippet below). One thing we should note is the awkwardness of having to bundle together the year restriction (2012-2019 → 2015-2018) with the 12 month restriction. Unfortunately, this table was generated in the data build process by Inovalon and we cannot go and disentangle it.

> *Between 2012-2019, about 1.3 million individuals were newly enrolled in the Medicare program at age 65, with 56% of them enrolled in MA. After our sample restrictions, our final sample composes of 205,557; 180,087 (87%) enrolling in FFS, and 25,470 (13%) enrolling in MA. Of the restrictions we impose, the ones that reduce our sample the most are a) the restriction to 2015-2018; b) the requirement for at least 12 months of coverage in the Inovalon data at age 64; c) no concurrent enrollment in commercial coverage at age 65; d) enrollment in a non-EGWP plan; and e) tracking of MA claims within the Inovalon dataset. In contrast, the requirements that enrollees stay in their plans continuously; that Inovalon tracks their commercial plans well; and that they enroll at age 65 do not bind as tightly. The final two requirements (that MA claims must be tracked, and that enrollees must be in non-EGWP plans) is especially binding for the MA component of our sample, resulting in MA underrepresentation within our sample; in contrast, we observe the universe of claims for FFS enrollees. We provide a step-by-step evolution of how our sample changes as we add restrictions in Appendix Table A1.*

c. This makes sense—we have updated this figure to now show the within-sample density instead, as you suggest.

d. Sorry that this was unclear. We now describe it in the main text:

> *First, we can compare MA enrollees in our analytic sample to MA enrollees who we can track at age 64, when they are enrolled in an Inovalon client ESHI plan, but not age 65, since they enroll in MA plans that are not Inovalon clients.*

and have made our description within the table more clear:

> *This table presents summary statistics for three cohorts of enrollees whom we continuously observe one year before and after age 65: enrollees who enrolled in Medicare Advantage plans captured by Inovalon (the first and third columns), enrollees who enrolled in Fee-For-Service at age 65 (the second and fourth columns),* ***en-***

> ***rollees who were enrolled in employer-sponsored plans captured by Inovalon, but then enrolled in Medicare Advantage plans that are not captured by Inovalon (the last column).***

(R4.2) *The authors correctly observed that even though the DiD design nets out any static differences in outcomes between treatment and control, there may still be threats to validity both from selection on future utilization, as well as from ESHI being a very heterogenous market (preventing an apples-to-apples comparison). To address these concerns, the authors use a matching strategy to look for treatment and control patients who look similar in terms of observables. In fact, Table 1 shows one dimension of such heterogeneity – plan type. Patients in ESHI HMO are more likely to later enroll in MA. Since HMOs are often paid by capitation, providers have stronger incentives to "upcode" patients enrolled in ESHI HMO plans, as opposed to those in POS or PPO plans. This type of upcoding behavior can complicate the interpretation of pre-Medicare differences as the "ground truth", as it affects the inputs (e.g. RAF and HCC) into the matching algorithm. To address this type of bias, the authors could consider an alternative specification where they do not match on health but instead match on demographics and other observables.*

We think this is less of an issue since we do all of our matching (and analysis) within plan type, both at age 64 and at age 65. Even when we study risk adjustment, our primary estimates use within cohort (state x urban/rural x plan type x quarter of birth) estimates, which should net out any major differences in coding across individuals at age 64.

(R4.3) *I would encourage the authors to dig deeper into the economic mechanisms that explain the fiscal costs results. On page 22, the authors write: "Though PPO plans do not appear to be favorably selected at all, they raise fiscal costs substantially. We estimate this largely comes from the fact that they (1) raise utilization somewhat, rather than lowering it; and (2) they submit high risk-adjusted bids given their underlying costs." While these are great accounting explanations, they don't quite get to the core economic mechanisms – what causes PPO plans to submit high bids? Is it a market power story, a demand story? I understand that this may be a significant undertake (and perhaps better suited for a separate paper) but given that the HMO-PPO divide is central to the message of the paper, it would be really helpful to understand the economics behind this phenomenon. The HMO-PPO divide is especially intriguing considering that whether to offer HMO or PPO is a choice of the insurer.*

We think doing this a lot of justice would be a big undertaking–we would likely have to estimate a "deep" model of plan bidding and conduct that we are not fully equipped to do. One thing we should note is that, in response to R1.4, we have changed our discussion of the "plan bidding" residual. We now discuss it in more detail up front:

> *When taking out all of these components, the final component reflects the residual gap between MA's private cost reductions and what it passes through to the government through the bidding system. This residual reflects three factors. First, plan bids may be greater than expected costs, to pay off their fixed costs or to earn markups when they face imperfect competition. Second, plans set bids uniformly across their entire enrollee population; the residual we estimate is only for 65-year-olds, but may be of different size or sign for other groups. Finally, any measurement error in estimating the other components will feed into this residual.*

And have changed our discussion of HMO/PPO differences to account for this:

> *Though PPO plans do not appear to be favorably selected at all, they raise fiscal costs substantially. We estimate this largely comes from the fact that PPOs (1) raise utilization somewhat, rather than lowering it; and (2) they have larger residual components. This latter effect suggests that PPOs have higher markups than HMOs, though this only applies to 65-year-olds. This might arise even if the two types don't have strictly different markups in a global sense if, for instance, PPOs enroll older populations and risk adjustment fails to fully correct for this, thus influencing plan bids (Orsini and Tebaldi, 2017).*

We realize this may not be a totally satisfying response to your comment, which is a good one.

(R4.4) *In Section 4.1, it would be helpful if the authors could clarify which numbers are taken from other sources, and which ones come from the analytical sample.*

Thank you, we have now done this. The new paragraph is below:

> *To compute current payments, we replicate MA's capitation formula. Current payments have four components: First, base payments, which are the product of the plan's bid (up to the county benchmark), and the risk scores of beneficiaries the plan enrolls. We extract bids from publicly-reported bidding data, and compute the risk scores from our claims data at age 64. Second, rebate payments. As described in Section 1, if plans bid below the county benchmark, they receive a share of the difference as rebates to spend on supplemental benefits, paid for by CMS, which we can compute using the bid, benchmark, and risk score. We construct these using public reporting on base rebate amounts combined with the risk score estimates in our data. Third, quality bonus payments. Plans that have high star ratings receive a greater share of the bid-benchmark gap as rebates, and also face higher benchmarks, allowing them to receive greater payments for the same bid. We reconstruct this from the publicly reported quality scores and bids, along with the risk score estimates in our data. Finally, we account for up-*

> *coding. We measure all risk scores at age 64, before beneficiaries enroll in Medicare. However, it is known that risk scores for MA enrollees are elevated due to more frequent coding, as well as due to tools like health risk assessments and chart reviews that enhance diagnostic documentation, but do not appear in claims data (Hammond et al., 2024; MedPAC, 2024). We calibrate a multiplicative net upcoding factor by taking the difference between estimates of upcoding from Geruso and Layton (2020) relative to the upcoding adjustment used in the MA plan payment formula. We describe each of these procedures in more detail in Appendix C.*

(R4.M1) *The exhibits in Appendix B seem out of order (Appendix Figures B1 through B4 are in between Appendix Tables B2 and B3).*

Thanks, these are now properly ordered.

References

Abelson, Reed and Margot Sanger-Katz, "Private Medicare Plans Misled Customers Into Signing Up, Senate Report Says," *New York Times*, 2022.

Agarwal, Rajender, John Connolly, Shweta Gupta, and Amol S. Navathe, "Comparing Medicare Advantage And Traditional Medicare: A Systematic Review," *Health Affairs*, 2021, *40* (6), 937–944.

Baicker, Katherine, Michael E. Chernew, and Jacob A. Robbins, "The Spillover Effects of Medicare Managed Care: Medicare Advantage and Hospital Utilization," *Journal of Health Economics*, 2013, *32* (6), 1289–1300.

Brown, Jason, Mark Duggan, Ilyana Kuziemko, and William Woolston, "How Does Risk Selection Respond to Risk Adjustment? New Evidence from the Medicare Advantage Program," *American Economic Review*, 2014, *104* (10), 3335–3364.

Cabral, Marika, Michael Geruso, and Neale Mahoney, "Do Larger Health Insurance Subsidies Benefit Patients or Producers? Evidence from Medicare Advantage," *American Economic Review*, 2018, *108* (8), 2048–2087.

Curto, Vilsa, Liran Einav, Amy Finkelstein, Jonathan Levin, and Jay Bhattacharya, "Health Care Spending and Utilization in Public and Private Medicare," *American Economic Journal: Applied Economics*, 2019, *11* (2), 302–332.

Duggan, Mark, Amanda Starc, and Boris Vabson, "Who Benefits When the Government Pays More? Pass-Through in the Medicare Advantage Program," *Journal of Public Economics*, 2016, *141*, 50–67.

_ , Jonathan Gruber, and Boris Vabson, "The Consequences of Health Care Privatization: Evidence from Medicare Advantage Exits," *American Economic Journal: Economic Policy*, 2018, *10* (1), 153–86.

Geruso, Michael and Timothy Layton, "Upcoding: Evidence from Medicare on Squishy Risk Adjustment," *Journal of Political Economy*, 2020, *128* (3), 984–1026.

Hammond, Stuart, Andy Johnson, and Luis Serna, "The Medicare Advantage program: Status report," 2024. https://www.medpac.gov/wp-content/uploads/2023/10/MedPAC-MA-status-report-Jan-2024.pdf.

Ippolito, Benedic, Erin Trish, and Boris Vabson, "Expected Out-Of-Pocket Costs: Comparing Medicare Advantage With Fee-For-Service Medicare," *Health Affairs*, 2024, *43* (11), 1502–1507.

Jacobson, Gretchen, Tricia Neuman, and Anthony Damico, "Do People Who Sign Up for Medicare Advantage Plans Have Lower Medicare Spending?," 2019. https://www.kff.org/medicare/issue-brief/do-people-who-sign-up-for-medicare-advantage-plans-have-lower-medicare-spending/.

Jung, Jeah, Caroline Carlin, and Roger Feldman, "Medicare Advantage Has Lower Resource Use and Better Quality of Care than Traditional Medicare," *American Journal of Health Economics*, forthcoming.

Lieberman, Steven M., Paul Ginsburg, and Samuel Valdez, "Medicare Advantage Enrolls Lower-Spending People, Leading to Large Overpayments," 2023. https://healthpolicy.usc.edu/research/ma-enrolls-lower-spending-people-leading-to-large-overpayments/.

MedPAC, "Estimating Medicare Advantage coding intensity and favorable selection," 2024. https://www.medpac.gov/wp-content/uploads/2024/03/Mar24_Ch13_MedPAC_Report_To_Congress_SEC.pdf.

_ , "The Medicare Advantage program: Status report," 2025. https://www.medpac.gov/wp-content/uploads/2025/03/Mar25_Ch11_MedPAC_Report_To_Congress_SEC.pdf.

Miller, Keaton, Amil Petrin, Robert Town, and Michael Chernew, "The Optimal Geographic Distribution of Managed Competition Subsidies," 2023.

Newhouse, Joseph P., J. Michael McWilliams, Mary Price, Jie Huang, Bruce Fireman, and John Hsu, "Do Medicare Advantage Plans Select Enrollees in Higher Margin Clinical Categories?," *Journal of Health Economics*, 2013, *32* (6), 1278–1288.

_ **, Mary Price, J. Michael McWilliams, John Hsu, and Thomas G. McGuire**, "How Much Favorable Selection Is Left in Medicare Advantage?," *American Journal of Health Economics*, 2015, *1* (1), 1–26.

Nicholas, Lauren Hersch, Dan Polsky, Michael Darden, Jianhui Xu, Kelly Anderson, and David J. Meyers, "Is there an advantage? Considerations for researchers studying the effects of the type of Medicare coverage," *Health Services Research*, 2024, *59* (1), e14264.

Ochieng, Nancy and Jeannie Fuglesten Biniek, "Beneficiary Experience, Affordability, Utilization, and Quality in Medicare Advantage and Traditional Medicare: A Review of the Literature," 2022. https://www.kff.org/report-section/beneficiary-experience-affordability-utilization-and-quality-in-medicare-advantage-and-traditio

Orsini, Joe and Pietro Tebaldi, "Regulated age-based pricing in subsidized health insurance: Evidence from the Affordable Care Act," 2017. BFI Health Economics Series 2017-02.

Schwartz, Aaron L., Atul Gupta, and Amol S. Navathe, "How Does Medicare Advantage Affect Health Care Use? Evidence from Beneficiary Transitions," 2023.

_ **, Khalil Zlaoui, Robin P. Foreman, Troyen A. Brennan, and Joseph P. Newhouse**, "Health Care Utilization and Spending in Medicare Advantage vs Traditional Medicare: A Difference-in-Differences Analysis," *JAMA Health Forum*, 2021, *2* (12), e214001.

Vatter, Benjamin, "Quality Disclosure and Regulation: Scoring Design in Medicare Advantage," 2024.

Wooldridge, Jeffrey M., "Distribution-Free Estimation of Some Nonlinear Panel Data Models," *Journal of Econometrics*, 1999, *90* (1), 77–97.

_ , "Simple approaches to nonlinear difference-in-differences with panel data," *The Econometrics Journal*, 2023, *26* (3), C31–C66.

Zahn, Matthew, "Entry and Competition in Insurance Markets: Evidence from Medicare Advantage," 2025.